INTERVALOS DE CONFIANZA Y PRUEBAS DE HIPOTESIS: UNA COMPARACION ENTRE DOS METODOS PARA ESTIMACION DE PARAMETROS

H. GIORGINI y A. GARSD[1]

Recibido: 16/09/98
Aceptado: 23/12/98

RESUMEN

Las dos formas más usuales de inferencia estadística son la estimación a través de intervalos de confianza y las pruebas de hipótesis. Ambas formas de inferencia permiten llegar a conclusiones similares. Sin embargo, se observa en el ambiente científico una sobrevaluación del alcance de las pruebas de hipótesis, mientras que los intervalos de confianza tienen la imagen de una herramienta rudimentaria. El valor *p*, resultante de una prueba de hipótesis, puede ser de gran utilidad cuando no se dispone de una estimación directa de un efecto o cuando dicha estimación es difícilmente interpretable. No obstante, es pobre la información que proporciona si se lo presenta aislado de sus implicaciones o limitado a las categorías de «significativo» o «no significativo». Los intervalos de confianza suministran, además de toda la información que provee una prueba de hipótesis, una rica información adicional. No existe razón para limitar la posibilidad de los resultados de un ensayo a una dicotomía cuando el intervalo de confianza presenta la diversidad de los valores reales y potenciales en estudio.

Palabras clave: Inferencia estadística, intervalos de confianza, pruebas de hipótesis, valor *p*.

CONFIDENCE INTERVALS AND HYPOTHESIS TESTING: A COMPARISON BETWEEN TWO METHODS OF PARAMETER ESTIMATION

SUMMARY

The two most common approaches to statistical inference are confidence intervals and hypothesis testing. Both forms of inference may lead to similar conclusions. Yet, in many a scientific context the emphasis seems to be placed on hypothesis testing, while the confidence interval is regarded as a rather more rudimentary tool. In fact, however, p-values from hypothesis testing are preferable only if no direct estimations are possible or interpretable. Otherwise, a p-value is less useful, particularly as a consequence of an isolated decision between two categories, such as significant or statistically non-significant results. By contrast, confidence intervals include information needed to test many candidate hypotheses. Thus, there is no valid reason to restrict results from an agricultural experiment to an oversimplified dichotomy, whenever confidence intervals can be used to explore specific as well as general, or potential, study outcomes.

Key words: Statistical inference, confidence intervals, hypothesis testing, p-value.

INTRODUCCION

Es usual que un ingeniero agrónomo se encuentre ante situaciones como la de advertir que un determinado fungicida utilizado en algunas parcelas de un campo ha sido más efectivo que otro fungicida empleado en distintas parcelas del mismo campo. Puede asimismo observar que un nuevo tipo de balanceado produce en algunos animales resultados más satisfactorios que el balanceado de uso habitual. Este profesional se preguntará si el fungicida en cuestión podría utilizarse en el resto del campo con idénticos resultados o si el nuevo balanceado será igualmente efectivo cuando se aplique a otros animales. La superioridad

[1]Cátedra de Estadística, Facultad de Agronomía, UBA. Av. San Martín 4453 (1417) Buenos Aires.

observada en ambos casos bien podría deberse al azar. La pregunta es entonces cómo saber si una diferencia observada persistiría en repetidos ensayos o si se trata de un mero producto de la casualidad.

Cuando el investigador se enfrenta a la necesidad de extraer conclusiones acerca del comportamiento de una variable en una población conociendo el comportamiento de dicha variable en una parte de esa población, el sentido común no siempre alcanza ni es aceptado por la comunidad científica como universalmente válido, por lo cual es imprescindible valerse de una metodología sistemática. Dicha metodología ha sido desarrollada por la disciplina estadística bajo la denominación de estadística inferencial. A diferencia de los razonamientos deductivos de las matemáticas puras, el razonamiento de la estadística inferencial es inductivo y se fundamenta en las leyes de la probabilidad, por ejemplo en la consideración de qué sucedería si determinado ensayo se repitiera muchas veces. Este enfoque, llamado frecuentista, está bien documentado en numerosos textos elementales (ver, por ejemplo, Steel y Torrie, 1989; Moore, 1985).

Hay básicamente dos formas de inferencia estadística. Una de ellas es la estimación, que puede ser de punto, en la que se utiliza un solo valor numérico para estimar el parámetro correspondiente, o de intervalo, que se compone de dos valores numéricos y el intervalo comprendido entre los mismos, asociado con un determinado grado de confianza que sostenga la inclusión en dicho intervalo del parámetro estimado. Así puede afirmarse informalmente, por ejemplo, que la altura media poblacional de cierto cultivo se encuentra entre 18 y 21 cm, con un 95 por ciento de confianza.

La otra forma de inferencia estadística es la prueba de hipótesis, que se utiliza cuando el objetivo del investigador no es estimar un parámetro, sino determinar si la evidencia empírica provista por una muestra es consistente, con un determinado grado de confianza, con una hipótesis acerca de la población que generó la muestra. Por ejemplo, ante una diferencia observada entre los promedios muestrales del aumento de peso de dos grupos de terneros sometidos a distintas dietas, puede sostenerse, con una confianza del 95 por ciento, que entre ambos tratamientos existen diferencias significativas, rechazando así la hipótesis de igualdad. Ambos tipos de inferencia se encuentran interconectados y presentan menos diferencias de las que a menudo se suponen, ya que pueden utilizarse para arribar a las mismas conclusiones (Blalock, 1986).

Frecuentemente se observa en el ambiente científico una particular sobrevaluación acerca del alcance de las pruebas de hipótesis cuando se requiere extraer conclusiones poblacionales a partir de resultados muestrales. Estas se perciben como los indicadores más refinados y sofisticados para juzgar el rigor metodológico de un análisis estadístico y por lo general no aparecen objeciones cuando, en la estadística aplicada, el resultado de una prueba de hipótesis se traduce sin mayor preámbulo en una decisión de rutina. Por el contrario, los intervalos de confianza tienen la imagen de una herramienta difusa y rudimentaria, cuya presencia queda limitada a los libros de texto de estadística elemental.

Sin embargo, a través del intervalo de confianza, puede conocerse toda la información brindada por una prueba de hipótesis y aún más. No existe razón para limitar la posibilidad de los resultados de un ensayo a una dicotomía cuando el intervalo de confianza presenta la diversidad de los valores reales y potenciales en estudio (Vardeman, 1992).

La causa de esta paradoja aparentemente radica en el carácter de los problemas que impulsaron el desarrollo de la estadística a principios de siglo, relacionados en general con la industria y con la agricultura (Rothman, 1988), ámbitos en los que la necesidad demanda a menudo una toma de decisión urgente entre dos o más alternativas, al servicio del sostenimiento del proceso de producción. El intervalo de confianza no pudo adaptarse sin dificultades a tales requerimientos o bien no logró imponerse a la sencillez de considerar un resultado dicotómico como el que resulta de observar, en una prueba de hipótesis, si el valor p es superior o inferior a la probabilidad de cometer el

error de tipo I (rechazar la hipótesis nula siendo ella verdadera) que se ha asumido. El concepto frecuentista de intervalo de confianza no es de comprensión trivial e involucra tecnicismos (Cox y Hinkley, 1974).

Aun aceptando tales explicaciones, el argumento no tiene ya razón de ser. Se trata por tanto en este informe de revalorizar la utilidad e importancia de la estimación de intervalo como forma de inferencia estadística, para lo cual deben analizarse su alcance y sus limitaciones.

MATERIALES Y METODOS

Se utilizaron ejemplos basados en casos reales, tomados de trabajos de consultoría realizados en la Cátedra. Los datos fueron modificados a fin de resaltar las falencias de los valores *p* cuando se trata de distribuciones de datos muy homogéneas o muy heterogéneas.

RESULTADOS Y DISCUSION

El procedimiento que debe emplearse para obtener una estimación de intervalo se inicia con la decisión acerca del riesgo de error que se está dispuesto a asumir al afirmar que el intervalo obtenido contiene al parámetro que se desea estimar. De este modo, si se asume, por ejemplo, un error del 5 por ciento, esto implica que por cada 100 intervalos que se obtengan con este procedimiento, 95 intervalos incluirán al parámetro que se desea estimar, mientras que los 5 intervalos restantes no lo harán. Se presenta por lo tanto la siguiente alternativa: o bien el verdadero parámetro se encuentra entre los límites del intervalo obtenido, o bien el azar ha determinado que el intervalo obtenido sea uno de los pocos (5 de cada 100) en los que eso no sucede.

Nótese que, como ocurre siempre en la inducción estadística, la confianza no se asienta en un resultado particular de la muestra, en este caso el intervalo obtenido, sino en el procedimiento que se emplea para arribar a dicho resultado. Por eso, no es correcto decir que hay una probabilidad del 95 por ciento de que el intervalo contenga al parámetro, ya que éste constituye un valor fijo cuya probabilidad de quedar en el interior de un intervalo determinado sólo puede ser uno o cero, debido a que el parámetro está o no está en el intervalo. Siendo así, no tiene sentido en este caso hablar de probabilidad.

Más allá de la apariencia de sofisticación que resulta del empleo de pruebas de hipótesis, normalmente se acepta que éstas sirven para apuntar la evidencia brindada por los datos en favor o en contra de una afirmación, mientras que los intervalos de confianza son apropiados cuando la meta es estimar un parámetro. Se admite asimismo que cuando se utiliza un intervalo es porque importa la posición aproximada del parámetro como un todo para interpretar un resultado, mientras que en la prueba de hipótesis el énfasis se pone en la precisa ubicación de uno de los límites de dicho intervalo. No obstante, ambas formas de inferencia permiten llegar a conclusiones similares, proporcionando los intervalos, además de toda la información que provee una prueba de hipótesis, una rica y en ocasiones fundamental información adicional.

Aun aceptando que las pruebas de hipótesis pueden ser de gran utilidad para el desarrollo de disciplinas vinculadas con la toma de decisiones perentorias (entre las cuales podrían incluirse algunas especialidades agronómicas), no es trivial el hecho de poder desconocer las implicaciones de estas decisiones. No debe ignorarse que, en una prueba de hipótesis, las categorías «significante» y «no significante» son absolutamente arbitrarias y que aun disponiendo de notables evidencias a favor de la presencia de un efecto, bien podría éste ser extremadamente pequeño. Significación estadística y significación práctica no son en modo alguno conceptos equivalentes: un bajísimo valor *p* indica una muy fuerte evidencia de la existencia de un efecto, lo que es muy distinto a afirmar que se posee evidencia de la existencia de un efecto muy fuerte. Una alta significación estadística no implica grandes diferencias. Particularmente cuando se dispone de muestras grandes, las pruebas de hipótesis son muy sensibles y detectan, sin dar lugar a dudas, dispersiones incluso muy pequeñas de la hipótesis nula (Fleiss, 1993).

Es frecuente en disciplinas tales como la biología que un valor *p* altamente significativo pueda

estar asociado con resultados irrelevantes. Considérese, por ejemplo, una pequeña, pero estadísticamente significativa, diferencia entre el grosor medio del tronco de dos grupos de árboles, uno de ellos sometido a un riego intensivo y sistemático y otro librado al comportamiento de las lluvias. A la inversa, un resultado no significativo podría ser, sin embargo, de gran importancia biológica. Supóngase, por caso, la existencia, surgida a través de la observación crítica de la masa de los datos experimentales, de algún mínimo indicio, estadísticamente insignificante, que haga presumir la presencia de efectos nocivos irreversibles para la salud en una vacuna que, de superar las pruebas con simios, será destinada al consumo humano masivo. No es entonces conveniente tomar decisiones importantes limitándose a observar el valor *p* sin haber realizado previamente un análisis descriptivo de los datos y sin conocer el rango de resultados compatible con los datos en estudio. Un intervalo de confianza, por ejemplo, para una diferencia de promedios podría incluir al cero, con lo cual una prueba de hipótesis daría cuenta de la no existencia de diferencias significativas y llevaría a aceptar la hipótesis nula. Disponiendo sólo del resultado de la prueba de hipótesis, el investigador ignora si esto es debido a la variabilidad de los datos, al reducido tamaño muestral o a ambas cosas. Esta situación es fácilmente salvable con sólo observar el intervalo de confianza obtenido, ya que éste sumariza los resultados en forma clara y sin ambigüedad.

Por otra parte, el empleo de intervalos de confianza no elimina la necesidad de arriesgar supuestos acerca de la naturaleza de la población y del método de muestreo utilizado. Básicamente, los supuestos en un problema de intervalo son los mismos que se requieren en cualquier prueba de hipótesis, excepto el de suponer un valor hipotético para el parámetro estimado. Sin embargo, si bien los intervalos de confianza, a diferencia de las pruebas de hipótesis se ven limitados a una sola probabilidad y aun cuando el objeto explícito de su utilización en una estimación está en indicar el grado de exactitud de ésta, los intervalos de confianza constituyen también pruebas implícitas de una vasta serie de hipótesis, ya que un intervalo de confianza equivale a una prueba virtual de todo valor posible del parámetro que pueda suponerse. Si se supusieran valores hipotéticos del parámetro que se situaran al interior del intervalo de confianza, no se descartarían dichas hipótesis al nivel de confianza considerado y si, en cambio, se supusieran valores del parámetro que quedaran al exterior del intervalo, se sabría que estas hipótesis se descartarían. Así, habiendo obtenido un intervalo de confianza, se puede decir a simple vista cuáles habrían sido los resultados de la verificación de cada hipótesis (Walker, 1993).

No debe pasarse por alto el hecho de que la suposición de un valor hipotético y más aún, en el caso de utilizarse una prueba de hipótesis, la reducción a una dicotomía del posible espectro de resultados supone una mediatización subjetiva del investigador, entre la observación objetiva de los datos y el posterior resultado objetivo que éstos proporcionan. El ejemplo más familiar de cómo el valor *p* depende de una acción subjetiva es la elección de pruebas estadísticas unilaterales o bilaterales. En ambos casos los datos son los mismos, pero el valor *p* se modifica. Si bien la subjetividad también interviene cuando para construir un intervalo de confianza el investigador debe optar por un determinado nivel de significación, mientras que en una prueba de hipótesis el valor *p* permite prescindir de esta elección, en este último caso, el proceso es más transparente, ya que siempre está disponible la observación crítica de los datos (Triola, 1992).

EJEMPLO 1

Se realizó en el establecimiento «La Alegría», ubicado a 17 km de la ciudad de Junín (provincia de Buenos Aires), un ensayo para comparar, durante la crianza del ternero, dos tipos de concentrado: uno de origen comercial y otro factible de ser preparado en el tambo sobre la base de granos producidos en el establecimiento. Se usaron por tratamiento 11 terneros Holando Argentino de ambos sexos nacidos en el establecimiento, los que se alojaron en forma individual en estacas. Al

Cuadro N° 1: Ganancia de peso de terneros (en kg) por el uso de dos tipos de concentrados

Dieta 1: Balanceado comercial						Dieta 2: Balanceado casero					
39,0	27,0	28,5	27,5	29,5	24,0	23,0	20,0	27,0	34,5	23,5	37,5
36,0	28,0	39,5	24,0	43,0		31,0	19,0	18,0	24,5	24,5	
Media: 31,45 kg						**Media: 25,68 kg**					

inicio del ensayo, el peso promedio de los animales fue de 56,2 kg y la edad promedio fue de un mes de vida.

Ambos tratamientos consistieron en 4 litros de leche por ternero en dos tomas diarias, fardo de alfalfa y de pradera suministrados a voluntad y agua. Los mismos se diferenciaron en el tipo de balanceado. La dieta 1 era un balanceado comercial dado a voluntad y la dieta 2 era un balanceado casero dado a voluntad. Durante el ensayo se aplicó el plan sanitario que normalmente se realiza en el establecimiento, consistente en aplicar desparasitante externo e interno y vacuna contra neumoenteritis. El ensayo se dividió en una primera etapa de acostumbramiento, cuya duración fue de 15 días, y una segunda etapa que fue el ensayo propiamente dicho, cuya duración fue de 1 mes.

Durante los primeros 3 meses de vida, los terneros permanecieron con sus madres. A medida que se destetaban, pasaban a una pradera de buena calidad, donde además se les suministraban 4 litros de leche diarios y balanceado comercial a voluntad. Cuando se reunió la cantidad de animales requeridos, se comenzó con el ensayo, durante el cual las estacas fueron cambiadas de lugar diariamente. El peso vivo de los terneros se determinó al inicio y al final del ensayo, considerándose como variable respuesta la ganancia de peso en kg (diferencia entre peso al final y peso al inicio del ensayo). Los datos obtenidos se observan en el Cuadro N° 1.

Se realizó una prueba *t* de Student para comparar las ganancias de peso resultantes de ambas dietas. La hipótesis de igualdad fue aceptada (p=0,0503), concluyéndose que no existían diferencias significativas entre las mismas. Se obtuvieron luego los intervalos de confianza bilaterales al 95 por ciento correspondientes a los promedios de ambos tratamientos, observándose una leve superposición, hecho que confirmó el resultado anterior.

Nótese que, en este caso, si se hubiera mantenido la mera afirmación de la no existencia de diferencias significativas entre las ganancias de peso producidas por las distintas dietas (el valor p obtenido fue de 0,0503), no habría podido advertirse lo que expresan los intervalos de confianza bilaterales al 95 por ciento, es decir promedios de 31,455 kg con límites de 28,563 kg y 34,346 kg para la dieta 1 y 25,682 kg con límites de 22,791 kg y 28,573 kg para la dieta 2. La observación de la diferencia eventual entre los 22,791 kg, límite inferior de la dieta 1, y los 34,346 kg, límite superior de la dieta 2, es más informativa que la aseveración lisa y llana de que no existen diferencias significativas entre las dietas. Sólo si el estudio tuviera la potencia estadística adecuada sería válida la conclusión de que no existe ningún efecto de importancia estadística. De lo contrario, lo más que se puede afirmar es que, con el tamaño de muestra con el que se trabajó, no hay evidencia suficiente para asegurar la existencia de diferencias significativas entre las dietas.

No obstante, aun contando con la potencia estadística adecuada, un valor p puede evidenciar un efecto de importancia estadística, pero no un efecto de importancia práctica, el que sólo podría manifestarse a través de la información descriptiva que proporciona la lectura cuidadosa de un intervalo de confianza.

EJEMPLO 2

Este ejemplo fue tomado de un ensayo que tenía por objetivo evaluar la demanda *per capita* de seis productos hortícolas: papa, zanahoria, cebolla, lechuga, tomate y pimiento en la localidad

Cuadro N° 2: Ventas de cebolla (en kg) en comercios minoristas para 2 estratos de la localidad de Ramos Mejía

Estrato 1: Precio del m² edificado: $ 900-1100						Estrato 2: Precio del m² edificado: $ 500-600				
175	177	175	177	176	175	175	176	177	176	176
176	175	176	175	175	176	177	176	178	176	177
176	176	176	174	175	177	176	177	177	177	
177	175	175	174	176	176	177	176	176	177	
174	175	175	176	175	177	177	176	177	176	
176	176	176	176	176	176	176	177	177	177	
176	176	177	175	174		175	175	176	176	
176	176	175	177	175		176	177	177	177	
Media: 176,441 kg						Media: 175,652 kg				

de Ramos Mejía, partido de La Matanza, provincia de Buenos Aires, a través de encuestas realizadas en comercios minoristas. Dicha localidad, según el censo de 1991, cuenta con una superficie de 11,9 km² y una población de 116.672 habitantes. Se determinaron dos estratos de acuerdo con el precio del metro cuadrado edificado según inmobiliarias de la zona. Dicho precio oscilaba entre u$s 900-1100 en el primer estrato y entre u$s 500-600 en el segundo estrato. Posteriormente se confeccionó un listado completo de los comercios minoristas de hortalizas de cada estrato. En el primer estrato se encontraron 46 comercios y en el segundo estrato, 34 comercios.

La encuesta duró aproximadamente una semana (del 20 al 27 de noviembre de 1992) y se consideraron las compras, los desperdicios y las ventas en kg que resultaban de la diferencia correspondiente para cada uno de los seis productos. Las variables consideradas alcanzaron de este modo un total de 21. A efectos del presente ejemplo, se comparan las ventas totales (compras menos desperdicios) de cebolla para ambos estratos. Los resultados obtenidos se presentan en el Cuadro N° 2.

Se realizó luego una prueba *t* de Student para comparar las ventas de cebolla en ambos estratos. En este caso se rechazó la hipótesis de igualdad de ambos promedios ($p<0,05$), concluyéndose que existían diferencias significativas entre los promedios correspondientes a cada estrato. Se obtuvieron finalmente los intervalos de confianza bilaterales al 95 por ciento para los promedios de ambos tratamientos, observándose que, efectivamente, no existía punto de contacto entre ambos.

El valor *p* inferior a 0,0001 no deja en este caso lugar a dudas. Sin embargo, si se observan los intervalos de confianza bilaterales al 95 por ciento, puede advertirse que las medias son de 176,441 kg, con límites de 176,250 kg y 176,632 kg, para el estrato 1 y de 175,652 kg, con límites de 175,488 kg y 175,816 kg, para el estrato 2. Si bien las diferencias estadísticamente significativas entre los estratos son inobjetables, es evidente que, conociendo el objetivo del estudio, dichas diferencias son teóricamente irrelevantes. Esta información no es proporcionada por el valor *p*, que tan sólo manifiesta la fuerte evidencia de la existencia de una diferencia, sin informar nada acerca de la magnitud de la misma.

CONCLUSIONES

Es indudable, aun para quienes sostienen la utilidad prioritaria de las pruebas de hipótesis en la estimación de parámetros, que éstas se han malinterpretado y se han utilizado indiscriminadamente: las asociaciones o diferencias estadísticamente significativas se han considerado, erróneamente, equivalentes a asociaciones o diferencias importantes. Es cierto, asimismo, que los valores *p* pueden ser de gran utilidad cuando no se dispone de una estimación directa de un efecto o cuando dicha estimación es difícilmente

interpretable. Sin embargo, los valores *p* no deben presentarse aislados de sus implicaciones teóricas y prácticas y mucho menos en la forma degradada de una calificación de «significativo» o «no significativo». Un agrónomo estudia y estima la magnitud de las relaciones biológicas, y la dicotomía producida por una afirmación infundamentada de significación podría estar, en ese contexto, fuera de lugar.

Los intervalos de confianza, por su parte, proporcionan estimaciones del espectro de verdaderas relaciones compatibles con un conjunto dado de observaciones. De esta manera, permiten conciliar resultados aparentemente contradictorios y generalmente introducen una advertencia apropiada de precaución en la interpretación de resultados «evidentes», ya que suelen decir más que lo que puntualmente se requiere de ellos.

Sin necesidad de clasificar a los resultados en categorías dicotómicas, hacerlo no puede ser más que contraproducente. Por otra parte, no hay regla estándar para comparar parámetros que tienen diferente «apoyo» del valor *p*. Esto hace particularmente difícil la interpretación de un efecto significativo pequeño en una muestra grande. La consecuencia práctica más conocida de que el valor *p* dependa sólo de una hipótesis es que un gran efecto en un estudio a pequeña escala y un efecto minúsculo en una investigación a gran escala generan un mismo valor *p*. Como el tamaño es parte esencial del carácter probatorio referente a la hipótesis de «no efecto», el valor *p* se hace inadecuado para cuantificar el carácter probatorio. El movimiento favorable a los intervalos de confianza representa un intento de considerar este tema, poniendo más atención en el tamaño del efecto.

Muchos de los inconvenientes ocasionados por la utilización de las pruebas de hipótesis son consecuencia del intento de darle sentido probatorio a índices que no están pensados para tal fin. Esto introduce la idea de que en los datos hay pruebas y verdades absolutas que pueden ser reveladas mediante técnicas estadísticas. A diferencia de los valores *p*, el uso de medidas de carácter probatorio fuerza a traer el juicio científico al análisis de los datos y muestra la diferencia entre lo que los datos dicen y lo que el investigador dice de los mismos (Goodman y Royall, 1993).

Las pruebas de hipótesis tienen una función de potencial utilidad en cada uno de los pasos importantes del análisis de los datos. Sin embargo, la utilización de las mismas como primer tratamiento de los datos, en lugar de un análisis informativo y descriptivo que permita evaluar consecuencias teóricas y prácticas, da lugar a malas interpretaciones de los resultados. Esta crítica no apunta a la utilidad de las pruebas de hipótesis en sí mismas, ya que son una simple herramienta de la que el investigador dispone, sino que intenta alertar sobre el uso abusivo y la interpretación errónea de las mismas.

BIBLIOGRAFIA

-**BLALOCK, H.** 1986. Estadística social, Sexta edición. Fondo de Cultura Económica. pp. 160-176, 211-221.

-**COX, D. and D. HINKLEY**, 1974. Theoretical statistics. Chapman and Hall. pp. 208-228.

-**FLEISS, J.** 1993. Las pruebas de significación tienen una función en la investigación epidemiológica: Respuesta a A. M. Walker en *Bol. Of Sanit. Panam.* 115 (2), pp. 155-159 (Traducción autorizada de «Significance tests have a role in epidemiologic research: reactions to A. M. Walker», Different Views, *American Journal of Public Health*, 1986. 76:556-558).

-**GOODMAN, S. y R. ROYALL**, 1993. Carácter probatorio e investigación científica en *Bol. Of Sanit. Panam.* 115 (3), pp. 235-249 (Traducción autorizada de «Evidence and scientific research» *American Journal of Public Health*, 1988; 78(12):1568-1574).

-**MOORE, D**. 1985. Statistics, concepts and controversies, Second edition. W. H. Freeman. pp. 296-304, 322-327.

-**ROTHMAN, K.** 1978, A show of confidence. *The New England Journal of Medicine.* 299:1362-1363.

-**STEEL, R. y J. TORRIE**, 1988. Bioestadística: principios y procedimientos, Segunda edición (primera en español). McGraw-Hill. pp. 59-63, 75, 83-88.

-**TRIOLA, M**. 1992. Elementary statistics, Fifth edition. Addison-Wesley. pp. 286-827.

-**VARDEMAN, S**. 1992, What about the other intervals?. The American Statistician. 46:193-197.-Walker, A. 1993. Cómo presentar los resultados de los estudios epidemiológicos en Bol. Of Sanit. Panam. 115 (2), pp. 148-154 (Traducción autorizada de «Reporting the results of epidemiologic studies», Different Views *American Journal of Public Health*, 1986. 76:556-558).